Keynote: Research Challenges in IR for Legal Discovery and Investigations

David D. Lewis
Brainspace, A Cyxtera Business
Dallas, TX USA
davelewis@daviddlewis.com

Abstract

The application of information retrieval (IR) technology to documents and data relevant to legal procedures (litigation, open government requests, antitrust reviews, etc.) has in the past two decades grown to become a multi-billion dollar industry. Many of the challenges faced by these e-discovery (electronic discovery) applications mirror those that have long been faced in investigations for corporate compliance, law enforcement, and national intelligence. These challenges are in some ways similar to, and in other ways very different from, those faced by enterprise search systems (another topic of this workshop).

Techniques for searching and mining billions of items spread across the planet have been the subject of intense research interest in IR. Techniques for actually finding something on your personal computer, much less discovering and synthesizing evidence from the computers (and phones and email stores) of ten thousand employees of an organization, have received oddly less research attention.

I make the case in this talk that IR research problems in e-discovery and investigations are every bit as intellectually interesting as those in web-scale IR. They are also much more accessible to researchers in academia, small companies, and other settings with modest resources.

I will present a list of such open research questions, solutions to which would have immediate practical and economic implications for e-discovery and investigations. These challenges are largely concerned with foundational issues in IR, including text representation, term weighting, statistical evaluation, and the basics of machine learning. No knowledge of latest flavor of distributed computing or nonconvex optimization is needed.

Bio

David D. Lewis, Ph.D. is Chief Data Scientist at Brainspace, A Cyxtera Business. He leads the data science team developing new information retrieval, machine learning, and natural language processing technologies for legal, investigatory, and intelligence applications. He is a Fellow of the American Society for the Advancement of Science, and won a Test of Time Award from SIGIR in 2017 for his 1994 paper introducing the uncertainty sampling algorithm for active learning.

In: Joint Proceedings of the First International Workshop on Professional Search (ProfS2018); the Second Workshop on Knowledge Graphs and Semantics for Text Retrieval, Analysis, and Understanding (KG4IR); and the International Workshop on Data Search (DATA:SEARCH18). Co-located with SIGIR 2018, Ann Arbor, Michigan, USA – 12 July 2018, published at http://ceur-ws.org